DOCUMENT RESUME

ED 340 737 TM 017 796

AUTHOR Halpin, Glennelle; McLean, James E.
TITLE Sources of Variability in the Angoff Standard-Setting Process.
PUB DATE Nov 91
NOTE 16p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (20th, Lexington, KY, November 12-15, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; Error of Measurement; *Evaluators; High Schools; *Interrater Reliability; Language Tests; Mathematics Tests; Reading Tests; *Scoring; *Secondary School Teachers; *Testing Problems; Test Reliability
IDENTIFIERS Alabama High School Graduation Examination; *Angoff Methods; *Standard Setting; Variability Measurement

ABSTRACT

Although the standard-setting method of W. H. Angoff (1971) has broad-based support in the research literature, inconsistencies in the resulting standards do occur. Sources of these inconsistencies are examined in a study of judges, competencies (items), rounds (replications), and the interactions among them. A modified Angoff approach was used to set standards on the Alabama High School Graduation Examination (AHSGE), Second Edition. Thirty-four subject matter high school teachers (judges) convened to set standards for the AHSGE. Data from round-two estimates if made, or round-one estimates otherwise, were analyzed using three-way factorial analysis of variance, with judges, rounds, and competencies as independent variables. The analysis was repeated for reading, language, and mathematics. For language and mathematics, judges accounted for the largest amount of variability. For reading, competencies were the largest source of variability. Although variability was the focus, the high degree of agreement found is noted, lending credence to the standards set despite some variability. Three tables present study data, and a 15-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED340737

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

GLENNELLE HALPIN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SOURCES OF VARIABILITY IN THE ANGOFF STANDARD-SETTING PROCESS

Glennelle Halpin

Auburn University

James E. McLean

The University of Alabama

Paper presented at the Annual Meeting of the

Mid-South Educational Research Association

Lexington, Kentucky

November 13-15, 1991

BEST COPY AVAILABLE

TM017796

SOURCES OF VARIABILITY IN THE ANGOFF STANDARD-SETTING PROCEDURE

Angoff (1971) proposed a method of setting standards of acceptable performance which is today among the most frequently used approaches. With the Angoff procedure judges predict the probability that each item on a test will be responded to correctly by minimally competent examinees (Jaeger, 1991). Cross, Impara, Frary, and Jaeger (1984) reported that judges using the Angoff approach expressed more confidence in the resulting standards than in the standards obtained with other standard-setting methods.

Halpin and Halpin (1987) determined the reliability and validity of the Angoff, Ebel, and Nedelsky methods of standard setting. They found that the internal consistency of the judges' ratings were high with the Angoff approach. The correlations between item ratings and the external criteria of actual item difficulty supported the validity of the Angoff procedure.

Although Angoff's standard-setting method has broad-base support in the research literature, inconsistencies in the resulting standards occur (Halpin & Halpin, 1981, 1987; Plake, Melican, & Mills, 1991; van der Linden, 1982). What are the sources of these inconsistencies? This study was designed to provide some answers to that question. Judges, competencies (items), rounds (replications), and the interactions among these sources were considered in a study in which a modified Angoff

approach was used to set standards on the Alabama High School Graduation Examination, Second Edition.

Method

Subject-matter teachers (N = 34) acting as judges were convened to set standards on the Alabama High School Graduation Examination using a modification of Angoff's technique. From over 35 schools throughout the state, they had been nominated by the system superintendent and selected by State Department of Education personnel. The criteria for the selection of the judges were (a) teaching at the high school level [9th through 12th grades]; (b) teaching the subject area they were to rate [reading, mathematics, or language]; and (c) teaching some students who were marginally competent in the respective subject areas. The teachers were chosen so as to represent city/county, urban/rural, and large/small school systems throughout the state. Both black and white males and females were included.

After being provided a general overview of the procedures employed in the development of the Alabama High School Graduation Examination, Second Edition, the judges were given an orientation to standard setting by one of the researchers. Standard setting was defined as establishing a cut score or point below which students in Alabama do not pass the Alabama High School Graduation Examination; establishing a minimum passing score on the Alabama High School Graduation Examination at or above which students do pass. The aim was to assure that passing the graduation examination is evidence of knowledge in each of the

subject areas measured. Judges were told that there are a number of approaches which may be used to set standards. Virtually all can be classified into one of three categories: theoretical, empirical, or judgmental (cf. McLean & Halpin, 1984). Each of these three strategies were briefly explained. Judges were informed that results from their judgmental approach would be considered along with standards set using the theoretical (Jensen, 1978) and empirical (Livingston & Zieky, 1982) approaches in the determination of the a standard to be recommended to the State Board of Education as the passing score on each of the three subtests of the Alabama High School Graduation Examination.

Following this overview, judges were given the following definition of a minimally competent student: one who can complete the required Carnegie units and has a minimum knowledge of [reading, language, or mathematics] to graduate from high school. They were then asked to visualize and describe (some orally) a minimally competent student they had taught. The researcher leading this discussion made sure that these descriptions included behavioral characteristics. Judges were reminded that minimally competent students are those just above the line that divides students into two categories: those who possess minimum knowledge to earn a high school degree and those who should not receive a diploma because of not possessing enough minimal knowledge.

According to their respective area of expertise, judges were then given the reading, language, or mathematics competencies

measured by the Alabama High School Graduation Examination along with associated test items. They were instructed to read the first competency and study the items designed to measure that competency. Then they were to estimate the number of similar items, out of 100, that a minimally competent student should be expected to answer correctly. They were to place that estimate next to that competency under the heading "Round 1" on their response sheets. The judges were then asked to discuss how they arrived at their estimates (but not what their estimates were).

After the researchers were satisfied that the judges understood the rating process, they asked the judges to repeat the rating process used with the first competency for each of the remaining competencies and associated test items. The researchers observed the rating process, which was executed independently, and remained available for questions. As each judge completed the task, the researchers checked to make sure that all estimates were between 0 and 100 and that none were missing.

When all judges had made their Round 1 estimates, they were provided with difficulty estimates (p values) for the items measuring each of the competencies in their subject area. They were then given an explanation of the p values with the caution that the values were computed using data from a pilot sample of students with a wide range of abilities and not just those who were minimally competent. Further, the judges were cautioned that the test was not a high stakes test for the pilot students.

The judges were then asked to reconsider their estimates. For the original estimates they wanted to alter, they put the new proportion by the appropriate competency under the column headed "Round 2" on their response forms. Judges were told that they did not have to change any of their estimates. They completed this second round of estimates, again working independently. When the judges completed the estimates for Round 2, they were thanked and dismissed. The researchers checked each judge's response form as it was turned in to be sure that estimates were provided for all competencies in Round 1.

Analyses and Results

Data from Round 2 estimates if made or Round 1 estimates otherwise were analyzed using three-way factorial analyses of variance with judges (random), rounds (random), and competencies (random) as independent variables. The analysis was repeated for reading, language, and mathematics. The primary reason for the analysis was to estimate the variance component for each potential source of variability. Estimates of the variance components are used to estimate the contribution of each source of variance to the total observed variance. The method used is similar to one suggested by Cronbach, Gleser, Nanda, and Jajaratnam (1972).

Results of the analyses of variance and estimates of the variance components for reading, language, and mathematics and shown in Tables 1, 2, and 3 respectively.

Insert Tables 1, 2, 3 about here

Discussion

The results provide some clues to the sources of variability in standards set using the Angoff procedure. For language and mathematics, judges accounted for the largest amount of variability (36.5% and 59.5% respectively) with the Judge x Competency interaction being the next most influential source of variance (30.2% and 19.6% respectively). The judges were all teachers who were trained in the use of the Angoff standard-setting procedure which should have reduced some of the variability in their estimates (Reid, 1991). However, they still differed in the standards they set. Less than perfect agreement among judges has been reported in other research (cf., Halpin & Halpin, 1987). Item characteristics--competencies, in this study--have been shown to influence standards set (Smith & Smith, 1988).

For reading, competencies were the largest source of variability accounting for 44.5% of the variance. The Competency x Judge interaction, accounting for 32% of the variance, was next. The reading competencies varied considerably in their level of complexity. Smith and Smith (1988) found that degree of inference required and vocabulary difficulty level affected standards set with the Angoff procedure.

An interesting finding was that rounds only accounted for an exceeding small amount of variability in the judges' estimates (1.4%, 0.0%, and .7% respectively for reading, language, and mathematics). Plake, Melican, and Mills (1991) averred that providing judges with performance data should decrease interjudge

variability. They contended that providing judges with the proportion of examinees who responded correctly to each item may serve as reality checks. Researchers (cf. Cross, Impara, Frary, and Jaeger, 1984; Norcini, Shea, & Kanya, 1988) have furnished judges difficulty estimates based on total groups of examinees with predictable results. That judges in the present study were told that the p values were based on the performance of all pilot students and not just those minimally competent could have caused the judges to discount the empirical data.

In an earlier study of the stability of judgments in an Angoff standard-setting procedure employed with the first edition of the Alabama High School Graduation Examination, McLean and Lockwood (1983) reported similar results. They found that rounds accounted for a minuscule amount of variance (.3%, .0%, and .0% for reading, language, and mathematics respectively). Judges (33.7%), competencies (22.4%), and Judge x Competency (36.4%) accounted for the majority of the variability in the estimates for reading. The Judge x Competency interaction accounted for 53.6% of the variance with the language test. Judges (40.0%) and Judge x Competency interaction (24.4%) accounted for the bulk of the variability in the mathematics estimates.

In the present study, we found that judges, competencies, and the interaction of these two variables explain a large part of the variability found in this standard-setting study using a modified Angoff approach. Although variability was the focus, the high degree of agreement found should be noted. Coefficient alpha reliability estimates for the judges in Round 1 were .82,

.95, and .98 for reading, language, and mathematics respectively. For Round 2 the alpha estimates were .85, .96, and .98 for reading, language, and mathematics respectively. Reliabilities such as these lend credence to the standards set even though some variability did exist.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.). Educational measurement (2nd ed.). Washington, DC: American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-129.

Halpin, G., & Halpin, G. (1981, November). Reliability and validity of the Ebel, Nedelsky, and Angoff methods of standard setting. Paper presented at the meeting of the Mid-South Educational Research Association, Little Rock, AR.

Halpin, G., & Halpin, G. (1987). An analysis of the reliability and validity of procedures for setting minimum competency standards. Educational and Psychological Measurement, 47, 977-983.

Jaeger, R. M. (1991). Selection of judges for standard-setting. Educational Measurement: Issues and Practice, 10(2), 3-6, 10, 14.

Jensen, O. (1978). Some logical procedures for setting passing points in the absence of direct criterion references. Princeton, NJ: Educational Testing Service. (Internal Document)

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

McLean, J. E., & Halpin, G. (1984, April). Setting passing scores for the Alabama High School Graduation Examination. Paper presented at the meeting of the American Educational Research Association/National Council on Measurement in Education, New Orleans.

McLean, J. E., & Lockwood, R. E. (1983, November). Stability of judgments in an Angoff-type standard setting procedure. Paper presented at the meeting of the Mid-South Educational Research Association, Nashville.

Norcini, J. J., Shea, J. A., & Kanya, D. T. (1988). The effect of various factors on standard setting. Journal of Educational Measurement, 25, 57-65.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. Educational Measurement: Issues and Practice, 10(2), 15-16, 22, 25-26.

Reid, J. B. (1991). Training judges to generate standard-setting data. Educational Measurement: Issues and Practice, 10(2), 11-14.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. Journal of Educational Measurement, 25, 259-274.

van der Linden, W. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 4, 295-305.

Table 1

Estimates of Variance Components for Reading

Source	SS	df	MS	Variance component	
				Estimate	Percentage
Judges	5188.13	11	471.65	8.41	8.10
Competencies	24303.83	20	1215.19	46.53	44.50
Rounds	453.34	1	453.34	1.51	1.40
J x C	17110.70	220	77.78	33.41	32.00
J x R	567.37	11	51.58	1.93	1.80
C x R	633.83	20	31.69	1.73	1.70
Error	2409.46	220	10.95	10.95	10.50
Total	50666.66	503		104.47	100.00

Table 2

Estimates of Variance Components for Language

Source	SS	df	MS	Variance component	
				Estimate	Percentage
Judges	24209.52	9	2689.95	54.59	36.50
Competencies	15330.29	22	696.83	28.63	19.10
Rounds	7.57	1	7.57	0.00	0.00
J x C	21117.93	198	106.66	45.13	30.20
J x R	796.67	9	88.52	3.14	2.10
C x R	748.08	22	34.00	1.76	1.20
Error	3247.18	198	16.40	16.40	10.90
Total	65457.24	459		149.65	100.00

Table 3

Estimates of Variance Components for Mathematics

Source	SS	df	MS	Variance component	
				Estimate	Percentage
Judges	63092.94	11	5735.72	112.80	59.50
Competencies	8109.46	23	352.58	9.23	4.90
Rounds	686.88	1	686.88	1.39	.70
J x C	22810.94	253	90.16	37.09	19.60
J x R	2717.89	11	247.08	9.63	5.10
C x R	1306.41	23	56.80	3.40	1.80
Error	4044.31	253	15.99	15.99	8.40
Total	102768.83	575		189.53	100.00